Draft Assembly Improvement

- **Motivation and background**
- **Goals and objectives**
- **Methods**
- **Results and discussion**
- **Still to do**

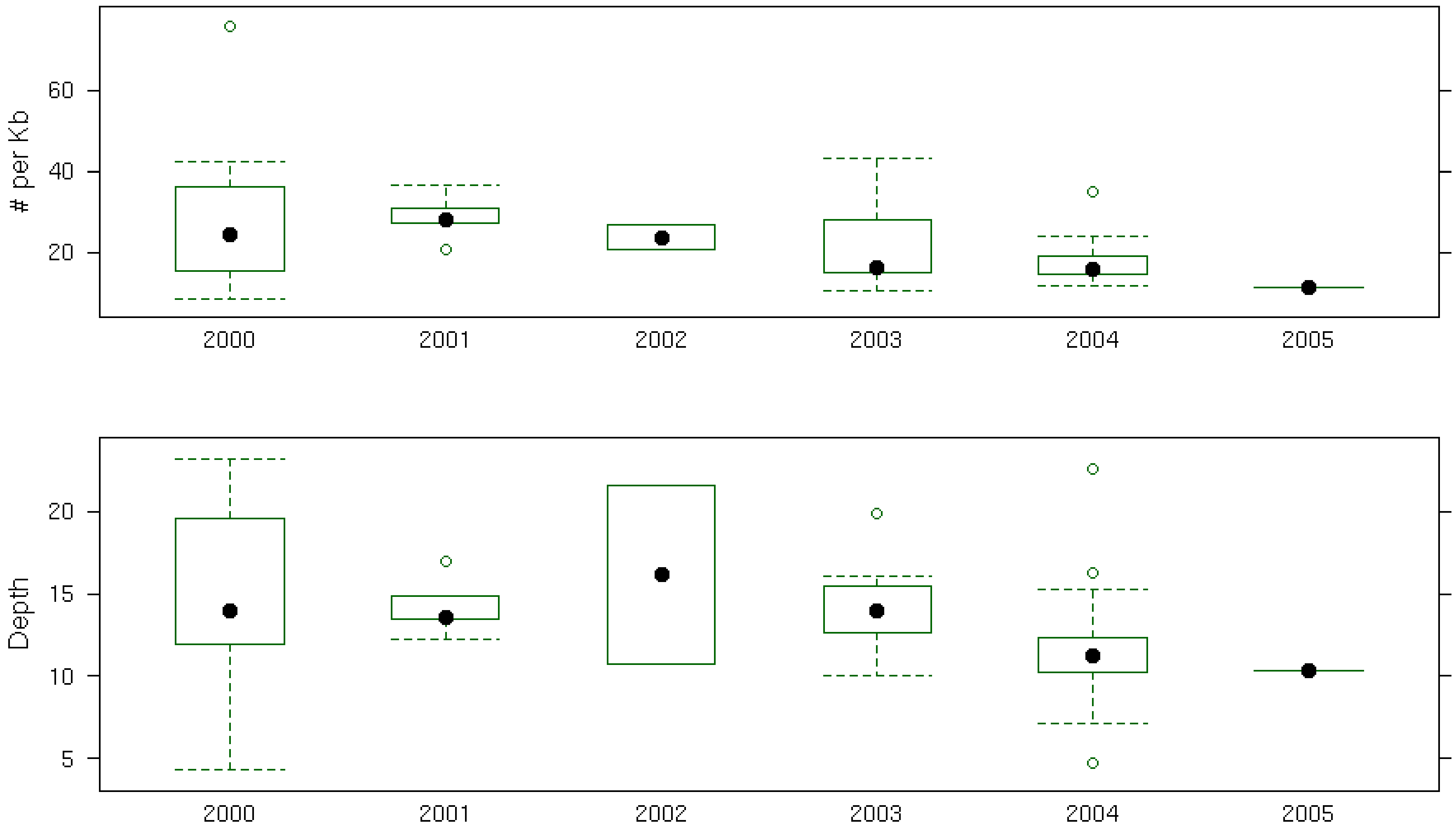
History

- **Early projects contaminated**
  - **Some assemblies failed to complete**
  - **Many long running assemblies**
  - **Lots of ad-hoc repairs and clean up**
- **Sequenced more reads than needed**

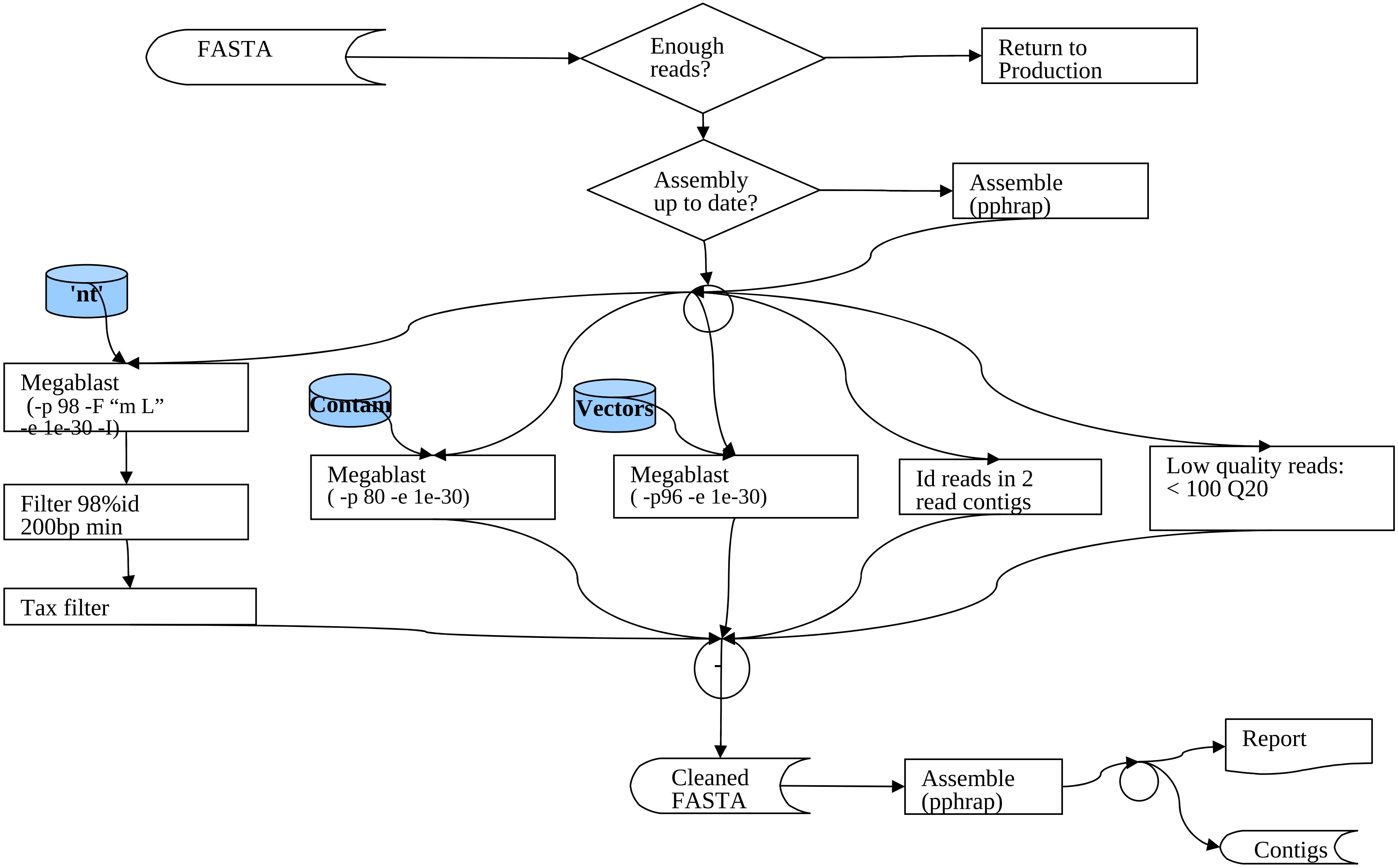
Goals

- **Do no harm**
- **Clean up obvious problems**
- **Make best use of existing data**
- **Create better assemblies for finishing**
- **Clean up public submissions**

Reads sequenced per 1000 bp of Genome



Flowchart



FASTA

Enough reads?

Return to Production

Assembly up to date?

Assemble (pphrap)

'nt'

Megablast (-p 98 -F “m L” -e 1e-30 -I)

Filter 98%id 200bp min

Tax filter

Contam

Megablast (-p 80 -e 1e-30)

Vectors

Megablast (-p96 -e 1e-30)

Id reads in 2 read contigs

Low quality reads: < 100 Q20

-

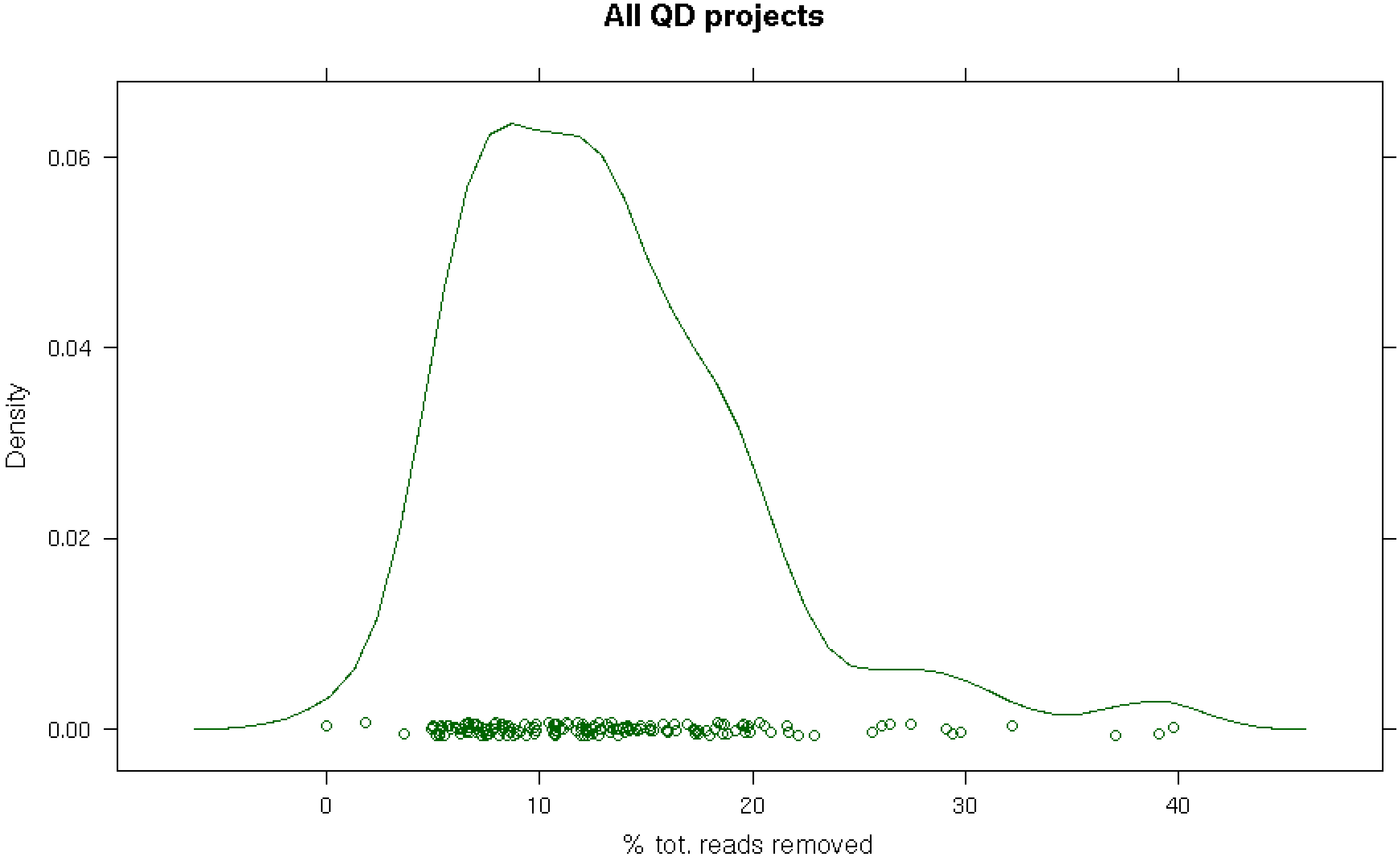
Cleaned FASTA

Assemble (pphrap)

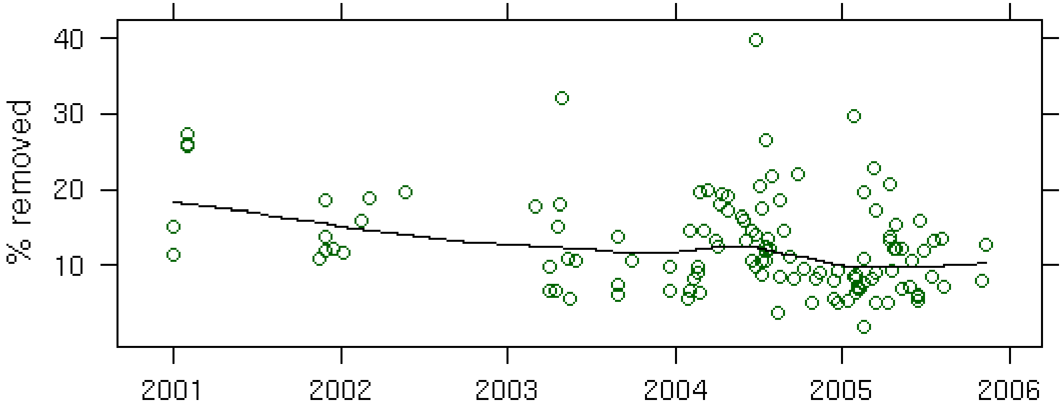
Report

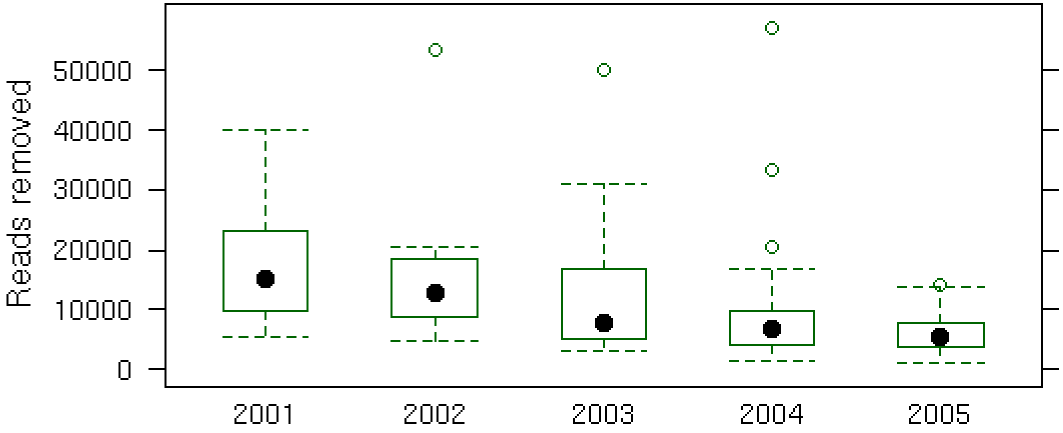
Contigs

What's not used



Process improvement ?



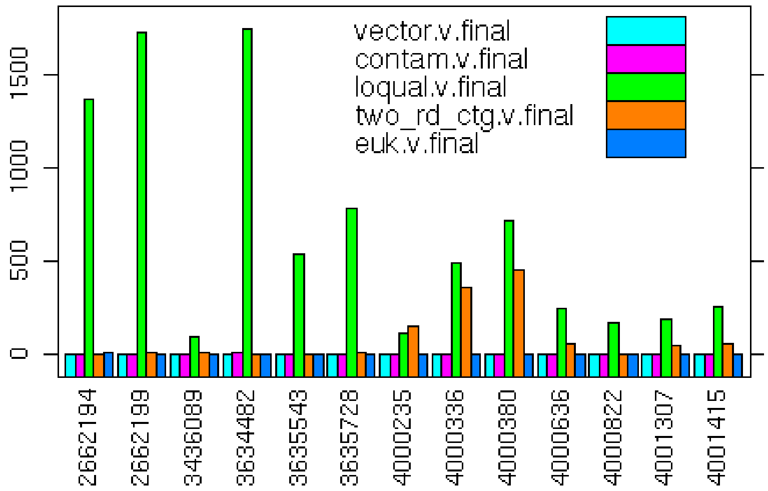


DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Breakdown of removed reads mapped to finished genomes

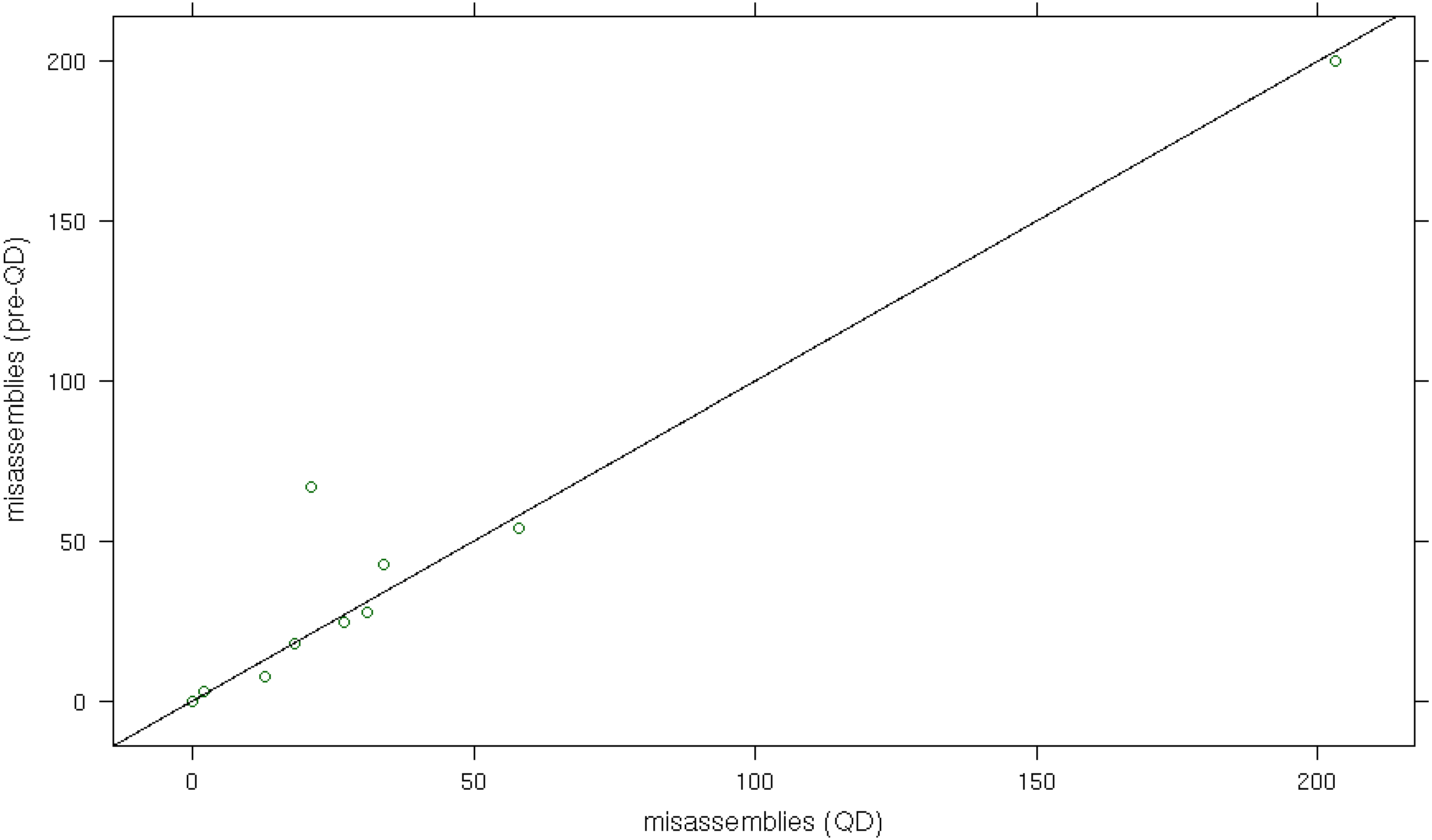
% removed with BLAST hits to finishe				Breakdown of removed reads				# removed reads hitting finished				
Low Quality	2 Rd Contig	Contam.	Euk.	Low Qual.	2 Rd Contig	Contam.	Euk.	loqual.v.final	2 Rd Ctg.v.final	contam.v.final	euk.v.final	vector.v.final
10.4	0.0	1.1	41.7	13146	5	280	12	1367	0	3	5	1
12.5	0.9	0.0	0.0	13832	436	324	98	1730	4	0	0	0
2.8	3.5	1/0	1/0	3353	144		0	94	5	0	0	0
14.8	1/0	1.7	0.7	11862	0	352	144	1750	0	6	1	0
16.4	13.6	1/0	1/0	3249	22	0	0	532	3	0	0	0
18.2	0.4	1/0	1/0	4274	1632	0	0	778	7	0	0	0
2.0	28.3	0.0	0.0	5349	519	2	3	107	147	0	0	0
6.8	62.4	0.0	1/0	7172	577	4	0	485	360	0	0	0
8.0	22.2	0.0	0.0	8955	2027	49	7	720	450	0	0	1
22.7	68.4	1/0	0.0	1071	76	0	20	243	52	0	0	0
25.8	1/0	1/0	0.0	647	0	0	1	167	0	0	0	0
19.4	91.1	0.0	1/0	943	45	2	0	183	41	0	0	0
27.5	73.0	0.0	1/0	936	74	2	0	257	54	0	0	0

Do removed reads belong?

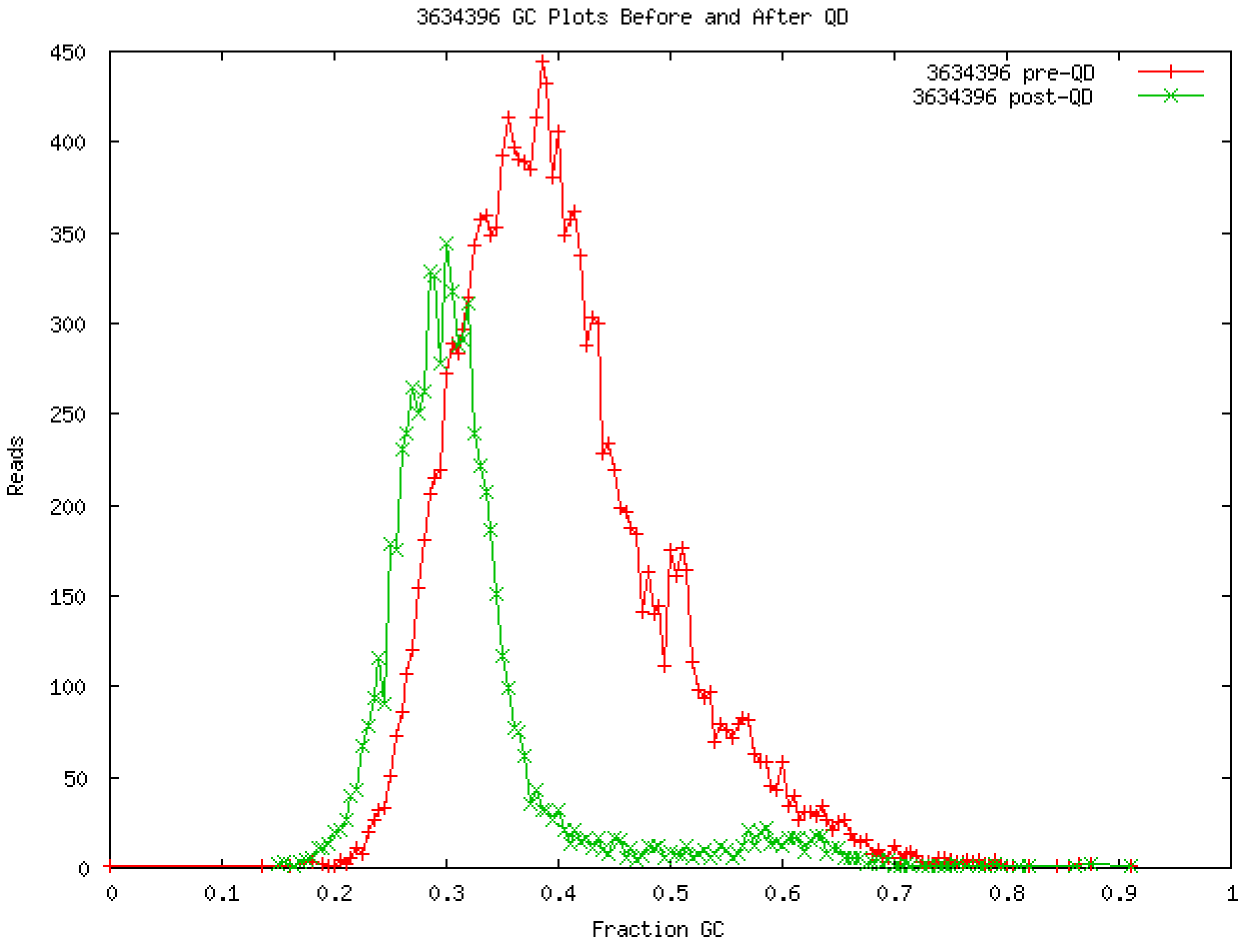


- Examined 14 finished projects
- BLAST removed reads against finished
- ~ 20% of low quality hit finished
- 2-read contigs are variable

Misassemblies



Benefits of QD



Summary

- **195 projects**
- **Projects appear to be improving**
- **Examined 14 finished projects – some number of removed reads belong in project**

Todo

- **Low quality filter could use tuning**
- **Better organism selectivity**

Acknowledgements

- Alla Lapidus
- Kerrie Barry
- Joel Martin
- Paul Richardson

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

LBNL-60077